DOCUMENT RESUME

ED 073 123 TM 002 366

AUTHOR	Frederiksen, Norman; Evans, Franklin R.
TITLE	Effects of Models of Creative Performance on Ability to Formulate Hypotheses.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	National Inst. of Child Health and Human Development (NIH), Bethesda, Md.
REPORT NO	ETS-RB-72-54
PUB DATE	Nov 72
NOTE	34p.; Draft
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Analysis of Covariance; Anxiety; Behavioral Objectives; College Students; Control Groups; *Creativity Tests; Experimental Groups; Higher Education; *Models; *Performance Tests; *Psychometrics; *Response Mode; Sex Differences; Tables (Data); Test Results; Tests; Training Techniques; Verbal Ability
IDENTIFIERS	*Formulating Hypotheses Test

ABSTRACT

The effects of sex, verbal ability, test anxiety, ideational fluency and training procedures on Formulating Hypotheses test performance were studied. Training consisted of presentation of models of "acceptable" responses that stressed either quantity or quality performance. Both the quantity and quality models were effective in modifying behavior in the expected direction. Ideational fluency was related to number of responses, and verbal ability was related to scores reflecting quality. Females were in general superior to males with respect to scores reflecting quantity of responses. Test anxiety was not significantly associated with performance. Weak evidence of treatment-anxiety and sex-vocabulary interactions was found. (Author)

ED 073123

RESEARCH BULLETIN

TM 002 066

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

RB-72-54

EFFECTS OF MODELS OF CREATIVE PERFORMANCE ON ABILITY TO FORMULATE HYPOTHESES

Norman Frederiksen

and

Franklin R. Evans

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the authors. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
November 1972

EFFECTS OF MODELS OF CREATIVE PERFORMANCE

ON ABILITY TO FORMULATE HYPOTHESES

Abstract

The effects of sex, verbal ability, test anxiety, ideational fluency and training procedures on Formulating Hypotheses test performance were studied. Training consisted of presentation of models of "acceptable" responses that stressed either quantity or quality of performance. Both the quantity and quality models were effective in modifying behavior in the expected direction. Ideational fluency was related to number of responses, and verbal ability was related to scores reflecting quality. Females were in general superior to males with respect to scores reflecting quantity of responses. Test anxiety was not significantly associated with performance. Weak evidence of treatment-anxiety and sex-vocabulary interactions was found.

EFFECTS OF MODELS OF CREATIVE PERFORMANCE ON ABILITY TO FORMULATE HYPOTHESES[1]

Norman Frederiksen and Franklin R. Evans
Educational Testing Service

In a previous study (Klein, Frederiksen, & Evans, 1969), the role of anxiety in learning a task requiring "creative" responses was investigated. The task involved the ability to think of hypotheses to account for research findings, using a test called Formulating Hypotheses (FH) (Frederiksen, 1959). The training consisted of presenting a model consisting of a list of "acceptable" responses after the completion of each item. The list contained 20 or more hypotheses, not all of which were necessarily of high quality. Students typically wrote only five or six responses; therefore the models were presumably perceived by Ss as emphasizing quantity more than quality of performance.

The specific hypothesis investigated was that the performance of anxious students would improve more than that of less-anxious students as a result of the training. The idea was that performance on such a task involves some self-censorship of ideas, especially by anxious people; in a situation calling for creative responses, people are likely to have more ideas than they report because they censor ideas that they think aren't "good enough" and therefore might lead to embarrassment. The feedback of models of "acceptable" responses, it was thought, would decrease the amount of censorship, thereby increasing the number of responses, and this effect would be greater for anxious than for less-anxious individuals.

It was found that the presentation of lists of model responses did produce a significant increase in the number of hypotheses written. It did not, however, produce an increase in the number of acceptable hypotheses written, and there was no evidence of transfer to another task requiring divergent production of semantic material (Guilford's Consequences test). Since the interaction of treatment and anxiety was not significant, the hypothesis that treatment would be more effective for anxious Ss was not confirmed. The overall results were tentatively interpreted as showing that the models comprising the experimental treatment were effective in changing the S's standards as to what constitutes satisfactory performance rather than his ability to deal with the problems.

An unexpected finding was that the relationship of anxiety to the number of hypotheses produced was curvilinear; poorest performance was associated with a middle level of anxiety and best performance with a low level of anxiety. This result is not predicted by drive theory and is contrary to Stennett's (1957) suggestion that the relationship between performance and drive level is nonmonotonic with the shape of an inverted U.

The purposes of the present investigation were (1) to attempt to replicate the results of the earlier study, particularly the effects of quantity models on number of responses and the curvilinear relationship of anxiety to performance; (2) to see if "quality" models of performance that were intended to improve quality of performance would be successful; (3) to investigate sex differences in creative performance; (4) to investigate the interactions of treatments, sex, ability, and anxiety in relation to quantity and quality of creative responses; and (5) to attempt to learn something about the processes involved in the improvement of performance.

Method

Subjects

The Ss in this study were 395 paid volunteers, mostly freshmen, at two state colleges in Pennsylvania. Approximately half were males and half females.

Measures

Formulating hypotheses. Five of the dependent measures were provided by a test called Formulating Hypotheses (FH), which is described more completely elsewhere (Frederiksen, 1959; Klein et al., 1969). Each item of the test consists of a graph or table showing findings from an actual research investigation in an area of social science. One item, for example, presents a graph showing that during a period following World War II employment decreased in mining and increased in all the other industries included in the survey; this finding is stated in words at the bottom of the graph. The S's task is to write hypotheses (possible explanations) that might account for, or help to account for, the finding. A sample item was presented showing examples of responses. Seven FH items were used instead of the eight in the earlier study.

In the previous study, two scores based on FH were used as dependent variables: Number of Hypotheses and Number of Acceptable Hypotheses, and they were also used in the present investigation. Number of Responses (FH Score 1) is the total number of nonduplicate hypotheses written by an S, and Number of Acceptable Responses (FH Score 2) is a subset of FH Score 1 responses that includes only hypotheses previously judged by a panel to be "acceptable." This score was used, in spite of its experimental dependence on FH Score 1, in order to make possible a more exact replication of

certain aspects of the previous study, and also because it reflects quality as well as quantity of ideas produced.

FH Score 3 is a new score called Average Judged Quality of the responses. It is the average of ratings made by two scorers (using a nine-point scale) of the quality of the responses, quality being defined in a sense that was consistent with the instructions to the examinees who took the FH test.

FH Score 4 is called Average Scale Value. It is another quality score, obtained by a method that makes it relatively independent of such qualities as length, handwriting, or grammatical correctness. The method made use of a master list of the hypotheses written by students, as derived from a content analysis. A panel of judges made evaluations of the hypotheses on this list, and a scale value was assigned to each response on the basis of these ratings. The scorer's task was merely to decide which listed hypothesis, if any, was similar to the one being scored, and to record the number of that listed hypothesis. A computer later assigned the corresponding scale value to the response.

FH Score 5 is Average Number of Words per response. It was included to provide some insight into the processes involved in modifying behavior by presenting models, and the cues used by scorers in evaluating responses.

An attempt was made to develop another score that would represent the rarity, unusualness, or originality of S's responses. The method involved getting a weighted score based on the number of ideas that occurred at frequencies below three specified levels. It was found that such scores were unreliable, even though frequencies as high as 40% were used as the basis for keying. Therefore no originality score was included in the study.

Each of the five FH scores was obtained separately for each of the seven FH items; this made it possible to obtain scores for subsets of items. Two subsets were used: (1) The first two items (which were completed before any feedback materials were presented), and (2) the last five items (which were all susceptible to influences of the experimental treatments). The five FH scores based on Items 1 and 2 are called FH Pretest Scores and were used as covariates in the statistical analysis. The five FH scores based on Items 3 to 7 were used as dependent variables. Unit weights were used in obtaining the pretest scores. The weights used in obtaining the five-item composite scores were their loadings on the first principal component resulting from a principal axes factor analysis of the intercorrelations of the five items. Five such factor analyses were done, one for each of the five FH scores. This method yielded a score representing the common variance in each of the one-factor systems. (However, since the loadings were about equal, unit weighting would have served about as well in this instance.)

Consequences test. Consequences is one of the tests used by Guilford (1967) to measure divergent production. Each item presents a hypothetical situation (e.g., "What would be the results if people no longer needed or wanted sleep?"), and the task is to list possible consequences of that situation. Two scores were obtained, using Guilford's scoring method: Consequences-Obvious and Consequences-Remote, representing Divergent Production of Semantic Units and Divergent Production of Semantic Transformations, respectively, in the structure-of-intellect model. These scores were used to measure transfer of experimental effects to another task involving divergent production and were treated as two more dependent variables.

Cognitive tests. Two tests from the Kit of Reference Tests for Cognitive Factors (French, Ekstrom, & Price, 1963) were also used: Advanced Vocabulary, a 36-item multiple-choice synonyms test; and the Theme Test, which requires S to write two themes, the score being merely the number of words written. These two tests measure the factors of verbal comprehension and ideational fluency, respectively, according to the French, Ekstrom, and Price manual. In the structure-of-intellect model, the Theme test represents Divergent Production of Semantic Units.

Test anxiety. The same inventory used in the previous study provided the Test Anxiety score; this scale contains items from Harleston's (1962) measure of test anxiety and the items from the Alpert-Haber (1960) debilitating anxiety scale. Since the Defensiveness scale showed little relationship to performance in the previous study, it was not included in this investigation.

Procedure

The procedure used was very similar to that employed in the previous experiment, except that there were two experimental treatments instead of one, and the data were obtained in one long evening session rather than in three separate sessions.

The experimental treatments consisted in providing models of acceptable performance, at the completion of each FH item (except the first), in the form of a list of hypotheses pertaining to that item. One treatment (the quantity treatment) was essentially the same as that used in the first study; it consisted of providing a fairly long list (18 to 26) of "acceptable hypotheses" illustrating ideas that S might have written in response to the preceding item. The other treatment (quality) was

similar except that each list of acceptable hypotheses included only the best ideas, carefully worded. The quality list typically included only six or seven hypotheses, and they were somewhat longer than those used for the quantity models.

Members of the control group received no models; instead, they were given various questionnaires to occupy their time in what appeared to be a relevant way.

The lists were intended to provide models of performance that Ss would try to emulate in one way or another. Both the quality and quantity materials are believed to show a rather striking contrast to the work of most students: The quantity lists contained many more responses than the average S wrote, while the quality lists were noticeably superior in quality, both with respect to ideas and wording. The materials (both quantity and quality) were represented to the Ss as responses written by college students that "were judged acceptable in that they give plausible explanation of the finding, although some...are no doubt better than others....These hypotheses are presented merely as examples of good responses by other students, in order to stimulate your thinking." In order to insure that the lists were read, S was instructed to study the list carefully, then to go back to his own list to make any revisions or additions he felt would improve the list. (The FH answer sheet produced a copy of S's responses. The original was removed before the feedback materials were presented, and the revisions were made on the copy. Only the original was used in scoring.)

All Ss from a given college were seated together in a large room, and all three treatment groups were handled simultaneously. Assignment of

treatments to Ss was accomplished by handing to every third S as he entered the room an envelope containing materials for one particular treatment. All documents were numbered as shown in Table 1, and instructions were given by referring to document numbers. The document in use at a particular period was taken by S from the top of the pile in his envelope and was placed at the bottom of the pile in the envelope when completed. Ss knew that the materials were not identical for all Ss, but they did not know the nature or purpose of the different treatments.

Insert Table 1 about here

Statistical Analysis

Means, standard deviations, and intercorrelations of all variables were computed for all Ss combined and separately for the three experimental groups. Similarly, means, standard deviations, and intercorrelations were computed for the seven FH items, once for each of the five FH scores. Reliabilities were computed where possible.

Because of the relatively large number of dependent measures, the method of analysis chosen was multivariate analysis of covariance (MANCOVA) (Clyde, Cramer, & Sherrin, 1966). There were seven dependent measures: the five FH scores based on Items 3 to 7, Consequences-Obvious, and Consequences-Remote. Since there were small differences between the two colleges in student performance, one covariate was the dichotomy college attended. Other covariates were the five FH Pretest Scores. These five scores are appropriately used as control variables whenever we are interested in change in performance on FH (i.e., when investigating effects of treatments).

The design factors were treatment, sex, vocabulary, ideational fluency, and test anxiety. However, it was not possible to use all five design factors in one analysis because some cell frequencies became too small. Instead, three separate MANCOVA's were done with overlapping factors. The three analyses employed the following design factors, with number of levels within each factor shown in parentheses:

1. Treatments (3), sex (2), ideational fluency (2), test anxiety (3)
2. Treatments (3), sex (2), ideational fluency (3), vocabulary (2)
3. Treatments (3), sex (2), vocabulary (2), test anxiety (3)

Three levels of test anxiety and of ideational fluency were used in order to make possible the detection of nonlinear relationships. Each of the three designs was used once with college attended as the only covariate, and once with the five FH pretest scores used as covariates in addition to college attended.

The MANCOVA model employed in the analysis first removes variance attributable to the four-way interaction, then lower-order interactions, and finally it deals with main effects. Each main effect was computed so that it was orthogonal to all other main effects and all interactions. Thus the R^2 may be interpreted as the percentage of variance uniquely attributable to the factor under consideration.

Results

Intercorrelations and Reliabilities

Table 2 presents the means, standard deviations, intercorrelations, and reliabilities of the variables used as covariates and as design variables, using data for all subjects combined. Since the treatments may

affect intercorrelations involving dependent variables, the intercorrelations of dependent variables are shown separately in Table 3 for the three treatment groups as well as for the total group; and in Table 4 are shown the correlations of the dependent variables with the other variables for the three treatment groups and for the total group.

Insert Tables 2, 3, and 4 about here

Reliabilities of the various measures are shown in the main diagonals of Tables 2 and 3. (Reliabilities are based on the total group.) The last five diagonal entries in Table 2 are the correlations between the two items making up the FH Pretests, corrected for double length. The first five diagonal entries in Table 3 are the reliabilities for the five-item FH test, computed by obtaining the average of the item intercorrelations and correcting for length by the Spearman-Brown formula. The correlations between the two-item and the five-item tests are shown in Table 4; these may be thought of as alternate form reliabilities. These correlations are, of course, attenuated by the fact that one of the two tests contains only two items. FH Score 5 (Average Number of Words) is the most reliable, and FH Score 4 (Average Scaled Value) is the least reliable of the scores. While the two-item pretest is adequate for use as a control variable, a longer test would obviously have been better. The data indicate that highly reliable measures of FH performance can be built by using a sufficient number of items.

Intercorrelations of FH test scores show that the FH Score 1 and FH Score 2 are highly correlated, as is to be expected since FH Score 2 is based on a subset of the responses contributing to FH Score 1. The

two scores designed to measure quality--FH Scores 3 and 4--are also highly correlated, in comparison with their reliabilities, and Score 3 has a substantial correlation with FH Score 2 (Number of Acceptable Hypotheses), which reflects quality as well as number of responses. The correlation between FH Score 3 (Average Rated Quality) and FH Score 5 (Average Number of Words) suggests either that raters tend to be impressed by long responses or that a more lengthy response is necessary for higher quality.

Since scores on the five-item FH scores are influenced by the experimental treatments, it is important to look at their intercorrelations and correlations with other variables separately for the three treatment groups. Treatments might influence the correlations as well as the means; if the relationships were altered appreciably, the meaning of a score, in terms of its factorial composition, could be changed, which might make interpretation of the MANCOVA results misleading.

Table 3 includes the intercorrelations of the dependent variables for the three treatment groups. There were some differences in correlations that might be attributable to treatments. For example, correlations involving FH Score 5 (Number of Words) were apparently reduced or made negative by the Quantity feedback. However, the differences are not great, and the pattern is generally the same for all three groups. The correlations with independent variables (Table 4) do not differ greatly for the three groups. It is therefore concluded that factor structure was not substantially altered by the treatments and MANCOVA results would, from this point of view, be interpretable.

Canonical Variates

Table 5 shows the correlations of the dependent variables with the four significant canonical variates (out of six computed) that were obtained

by using college attended and the five FH Pretest Scores as covariates, and without including any of the design factors. The canonical correlations (shown in the row labelled R) are correlations between the best-weighted combination of dependent variables and the best-weighted combination of covariates. These canonical variates are orthogonal, and they result from a step-down model (variance attributable to the first canonical variate is removed before computing the second, etc.). The correlations of dependent variables with the canonical variate may be used like factor loadings to interpret the canonical variates. (There are three sets of zero-order correlations and three R's for each canonical variate, which result from the analyses of the three different combinations of design factors described earlier. The reason for the slight differences among sets is that the pooled within-cell sums of squares differ slightly from one combination of design factors to another.)

Insert Table 5 about here

The first canonical variate was obviously defined by FH Score 5 (Average Number of Words). The canonical correlation between covariates and the dependent variables was high, about .75, no doubt in part because of the higher reliability of FH Score 5, which correlates about .98 with the canonical variate. The only other dependent variable with appreciable correlations was FH Score 3, the Average Quality Rating. These correlations suggest that there may be a tendency for raters to give higher ratings to the longer responses; or, possibly, longer responses are necessary for statements of high quality.

Canonical Variate II was defined mainly by FH Scores 1 and 2 (Number of Hypotheses and Number of Acceptable Hypotheses). Sizable correlations

also occurred for Consequences-Remote, but not for Consequences-Obvious, which suggests that the quantity of production on the FH test involves divergent production of semantic transformations rather than semantic units.

Canonical Variate III was mainly correlated with FH Score 3, the Average Quality Rating. Other substantial positive correlations were found for FH Score 2 (Number of Acceptable Responses) and FH Score 4 (Average Scale Value) reflecting the quality component in both these scores. In contrast to Canonical Variate II, the correlations with Consequences-Remote were negative, suggesting that the FH quality scores were quite different from the quantity scores with respect to the influence of fluency.

The last canonical variate, which was barely significant, was correlated mainly with FH Score 4, the Average Scale Value. FH Score 4 is the least reliable FH score, and a good deal of the variance attributable to it had already been allocated to Canonical Variate III. Thus there appear to be three major components in the domain of the dependent variables: length of responses, number of responses, and quality of responses.

When college attended was used as the only covariate, the canonical variate was significant ($p < .01$), showing that college attended was significantly related to the dependent measures; but the canonical correlation was only about .25. The correlations with the canonical variate may be interpreted as showing that one college was superior with regard to the quality score (FH Score 3), and the other was superior on Number of Hypotheses (FH Score 1) and on Consequences-Remote. No hypothesis can be offered to account for these differences other than the possibility of some unintended difference in the conditions under which the tests were administered. The results do justify our use of college attended as a covariate.

Multivariate Analysis

Tables 6, 7, and 8 present the salient findings of six MANCOVAs. The analyses differed with respect to the combination of design factors employed, as was described previously, and with respect to the covariates used (college attended only or college attended and the five FH Pretest scores). Table 6 shows the results for the design factors treatment, sex, ideational fluency, and test anxiety; Table 7 the results for treatment, sex, ideational fluency, and vocabulary; and Table 8 the results for treatment, sex, vocabulary, and test anxiety. For treatment and for interactions involving treatment, the results are reported only for the analyses where FH Pretest scores are used as covariates. For the remaining design factors and interactions, results are reported for analyses where college attended is the only covariate. The FH Pretest scores are used as covariates for treatment effects because we wish to use a measure of change in evaluating the effects of the quality and quantity models. Results for all the main effects are reported, but results for interactions are reported only if the significance level reaches the 5% level for either multivariate or univariate tests. The results shown for the main effects were computed in such a way that R^2 can be interpreted as the percentage of variance uniquely attributable

---

Insert Table 6 about here

---

to a particular factor.

A word about the contents of the three tables: The first column indicates what main effect (or interaction) is described in the corresponding row of the table. The second column shows the overall multivariate F-ratio

and its related p and R values. (The second canonical variate was not significant in any instance.) Except in the case of an interaction, mean scores on the first canonical variate for the appropriate subgroups are shown in the next column. (The grand mean is set at zero.) In the next column are shown the salient correlations of dependent variables with the first canonical variate; this information makes clear what constitutes each canonical variate and thus which dependent variables contribute most to the means shown in the preceding column. In the last column are shown the p-values for univariate tests for the dependent variables shown in the preceding column. An entry was made in the last column if (1) the R with the canonical variate was $\geq .30$; or (2) the univariate significance level was $\leq .05$.

Results for treatment, sex, ideational fluency, and test anxiety. The largest R in Table 6 is .46, for ideational fluency, whose relationship to the first canonical variate was highly significant ($p < .001$). The correlations of dependent variables with the canonical variate show that ideational fluency was related mainly to Consequences-Obvious, but also to Consequences-Remote and FH Scores 1 and (to some extent) 2. Univariate tests were significant for all these variables. The results appear to support Guilford's placement of both the Theme test and Consequences-Obvious in the same cell of the structure-of-intellect model, and they also show that the Number of Hypotheses score has a large component of divergent production of semantic units.

The second-largest R is .37, for treatments. The five FH pretest scores as well as college attended were used as covariates for this design factor. (The effects of treatments were greater--$R = .37$ as compared with .31--and

more clearly focused when these six covariates were used than when only college attended was used as a covariate, while the pattern of performance remained the same.) Quantity treatment improved performance as measured by FH Score 1 (Number of Hypotheses) and FH Score 2 (Number of Acceptable Hypotheses), while quality treatment tended to produce fewer, longer, and somewhat better responses. The univariate tests for the four FH scores were all significant, including those for FH Scores 3 and 5. A one-way analysis of variance, with FH Score 3 as the dependent variable and with college attended and FH Pretest Score 3 as covariates, showed that mean performance of the quality treatment group on FH Score 3 was significantly higher than that of the control group ($F = 7.34$; $p < .007$). Thus both quality and quantity treatments produced the expected changes in performance.

The R for sex is .26. Females were superior to males on a canonical variate that is positively correlated with Consequences-Obvious and FH Scores 2 and 5, and negatively correlated with Consequences-Remote. Thus females were found to be superior with regard to performance on tests that reflect number of hypotheses, number of words, and number of obvious consequences; but they were poorer on remote consequences.

The multivariate F for test anxiety was not significant ($p < .12$). The means show that low anxiety Ss tended to be superior on a canonical variate that correlates most highly with Consequences-Remote and FH Score 2 (Number of Acceptable Hypotheses). The univariate test for Consequences-Remote was significant ($p < .033$). Thus high anxiety appears to have suppressed performance on an ability similar to divergent production of semantic transformations. The means on the canonical variate for low, middle, and high anxiety groups showed no evidence of a nonlinear relationship such as was found previously.

The MANOVA test for a treatment-anxiety interaction yielded a nonsignificant F. However, the univariate test for FH Score 4 was significant ($p < .014$). Examination of the nine cell means for FH Score 4 shows a tendency for low-anxious Ss in the treatment groups to out-perform low-anxious Ss in the control group, while the opposite was true for high- and middle-anxious Ss. Thus either quantity or quality treatment tends to benefit low-anxious Ss more than middle- or high-anxious Ss.

Results for treatment, sex, ideational fluency, and vocabulary. Table 7 reports the MANCOVA results for four design variables that include vocabulary. Verbal ability, as measured by the Vocabulary test score, was significantly related ($R = .30$, $p < .001$) to a canonical variate that clearly reflects quality of performance on FH and a tendency to write long responses. The univariate tests were all significant, and the relationship to verbal ability was positive, as is shown by the means.

Insert Table 7 about here

The results for treatments, sex, and ideational fluency were all quite similar to those shown in Table 6, except that in the case of sex the univariate tests were not significant for FH Scores 2 and 5. Inclusion of vocabulary as a design variable has apparently removed some of the variance that was attributed to sex in the analyses reported in Table 6.

Univariate tests show some evidence of an interaction of sex and vocabulary involving FH Scores 2, 3, and 4, the measures reflecting quality of performance on FH. The means on FH Score 2 in the 2 x 2 interaction table show that females generally earned higher scores, but the difference between males of high and low verbal ability was much greater than that between females of high and low ability.

Results for treatment, sex, vocabulary, and test anxiety. Table 8 presents results for the third combination of design variables, which provides an opportunity to see if there is a vocabulary-anxiety interaction. None was found. The sex-vocabulary interaction did appear again, and in this analysis the multivariate test was significant ($p < .016$) as well as the univariate tests. The univariate test for the treatment-anxiety interaction was again significant for FH Score 4. Generally speaking, results for main effects were very similar to those found in the other two analyses except for some differences in details of the canonical variate for sex that are attributable to the variations in design factors.

Insert Table 8 about here

Means and Intercorrelations of FH Items

The rate of change in performance on the seven FH items is of interest because it might provide a basis for inferences about the processes involved in learning to formulate hypotheses under the experimental conditions. Two competing hypotheses are that (1) improvement reflects a change in ability, and (2) improvement reflects a change in standards as to what constitutes satisfactory performance. The first hypothesis implies a gradual process of learning, which would be reflected in a gradually rising curve; and the second would be more consistent with a sudden increase in performance following the first experience with one of the models. Although the shape of the curve could not rigorously demonstrate either process, evidence of either gradual or sudden improvement would make the corresponding hypothesis a bit more attractive.

Another possible approach, one involving individual differences, is to examine the intercorrelations of the item scores to see if treatment groups

differ from the control group with respect to the pattern of intercorrelations. Learning data are usually characterized by a simplex pattern (high correlations for adjacent trials and gradual reduction in correlation between trials as they become more widely spearated in time). If, on the other hand, the change in performance is sudden rather than gradual, one might expect low correlations of Items 1 and 2 with the remaining items, and high correlations between Items 1 and 2 and among Items 3-7.

Both of these approaches were tried; it was concluded that the data are too unreliable at the item level to yield interpretable findings.

Discussion

The effects of treatments on change in performance were shown to be highly significant; the proportion of variance in the canonical variate accounted for by treatments is about .13, when pretest scores are used as covariates. The effect of quantity treatment was basically to increase the number of FH responses and decrease the average number of words per response, while the effect of the quality treatment was to increase the average number of words per response, increase the quality of responses, and decrease the number of responses. The result of the earlier study is thus confirmed in that quantity models were found to increase quantity of performance. The effect of the quality treatment is smaller as judged by correlations of FH scores with the canonical variate, but a separate one-way analysis of variance confirms the finding that the quality models result in responses of higher quality. The change in performance tends toward a literal copy of the models in terms of number, length, and quality of responses. No evidence of transfer of training to the Consequences test was found.

Ideational fluency was found to account for a relatively large proportion of the variance in the domain of dependent variables ($\underline{R}^2 \cong .23$); and the other ability measure employed, the Vocabulary test, was also significantly related to performance ($\underline{R}^2 \cong .09$). These two ability measures predict quite different aspects of performance: ideational fluency is related to quantity of production (especially the Consequences test scores and FH Score 1), while vocabulary is related to quality of performance. A definition of creativity in terms of fluency would appear to be correct only if the quality of creative performance were ignored.

Sex was also found to be a significant factor ($\underline{R}^2 \cong .07$), although the proportion of variance contributed depends somewhat on what other design factors are included in the analysis (since sex is significantly correlated with anxiety, vocabulary, and, especially, ideational fluency). Females were generally superior on a canonical variate that correlates positively with Consequences-Obvious and FH Scores 2 and 5.

Test anxiety accounted for only a small amount of variance ($\underline{R}^2 \cong .04$); the canonical variate primarily reflected the Consequences-Remote score. High anxiety was associated with poorer performance, and there was no evidence of a nonlinear relationship.

A salient finding of the earlier study was a U-shaped relationship between test anxiety and Number of Hypotheses. A possible reason for the failure to replicate the curvilinear relationship is that the relationship of performance to anxiety is different for males and females (the previous study used only male subjects). Although a significant anxiety-sex interaction was not found, a plot (not shown) of FH Score 1 (the variable involved in the earlier study) against the three levels of test anxiety showed that

the same U-shaped curve existed for male Ss, while for females the curve was more or less linear and descending (high anxiety associated with fewer hypotheses). Thus the data at least suggest that the relationships for males are basically similar to those found previously and that sex differences exist.

The anxiety-verbal ability interaction found in the other study was also not replicated.

Weak evidence of a treatment-anxiety interaction was found involving FH Score 4. Since the hypothesis that motivated the original study was that anxious individuals would profit more from the treatments than less-anxious people, and that hypothesis was not then confirmed, the finding of even a weak treatment-anxiety interaction is of interest--even though FH Score 4 was involved rather than FH Score 1. A plot (not shown) of treatments against FH Score 4 means for the three levels of test anxiety showed relationships completely unlike those predicted for FH Score 1. The plot shows that for the control group, quality of performance was poorest for low-anxiety individuals, while for both treatment groups performance was higher for low-anxiety Ss. At higher levels of anxiety, the performance of control group members was superior to the treatment groups. Thus the relationship was the opposite of what had been predicted for FH Score 1. However, since FH Score 1 is a quantity score and FH Score 4 a quality score, the results may not be inconsistent with the original hypothesis.

A weak sex-vocabulary interaction was also found, the correlates of the canonical variate including FH Scores 2, 3, and 4, and Consequences-Remote. The first three of these variables emphasize quality of performance on FH rather than pure fluency. Examination of appropriate plots showed that

performance of females was generally superior, but there was relatively little difference between males and females of high ability while low-ability females were much superior to low-ability males.

Formulating Hypotheses appears to possess appropriate psychometric properties for further explorations in the realm of creative performance. It possesses a certain amount of face validity, the items being concerned with interpretation of real data obtained in various kinds of scientific undertakings, and therefore may possess certain advantages over such tests as "brick uses" and "consequences." The scores so far developed are reasonably adequate from the standpoint of reliability, and the interitem correlations tend to be sufficiently high that one could build a test of almost any reliability he desires by increasing the number of items. The span of abilities covered by the present five scores appears to include quantity of performance, quality of performance, and length of responses. It would be desirable to add scores measuring rarity or originality of responses. The study provides some evidence of the construct validity of the test, since the scores generally relate to other measures and to treatments in ways that are logical or in accordance with theoretical expectations. The use of tests like FH may be useful as provisional criterion measures in investigations of scientific creativity--its trainability, the influences of situational factors, and the cognitive, attitudinal, and temperamental characteristics associated with it.

References

Alpert, R., & Haber, R. M. Anxiety in academic achievement situations. Journal of Abnormal and Social Psychology, 1960, 61, 207-215.

Clyde, D. J., Cramer, E. M., & Sherrin, R. J. Multivariate statistical programs. Coral Gables, Fla.: Biometric Laboratory of the University of Miami, 1966.

Frederiksen, N. Development of the test "Formulating Hypotheses": A progress report. ONR Technical Report, Contract Nonr-2338(00). Princeton, N. J.: Educational Testing Service, June 1959.

French, J. W., Ekstrom, R. B., & Price, L. A. Manual for Kit of Reference Tests for Cognitive Factors. Princeton, N. J.: Educational Testing Service, 1963.

Guilford, J. P. The nature of human intelligence. New York: McGraw-Hill, 1967.

Harleston, B. W. Test anxiety and performance in problem-solving situations. Journal of Personality, 1962, 30, 557-573.

Klein, S. P., Frederiksen, N., & Evans, F. P. Anxiety and learning to formulate hypotheses. Journal of Educational Psychology, 1969, 69, 465-475.

Stennett, R. G. The relationship of performance level to level of arousal. Journal of Experimental Psychology, 1957, 54, 54-61.

Footnote

[1]This research was supported by the National Institute for Child Health and Human Development under Research Grant 5 PO1 HD01762. The authors wish to thank Albert E. Beaton and Charles E. Hall for their help with data analysis and Donald B. Rubin and William C. Ward for their critical review of the manuscript.

Table 1

Sequence of Presentations

Document Number	Group			Time (in minutes)
	Control	Quality	Quantity	
1		Personality Inventory		untimed
2		Advanced Vocabulary Test		8
3		Theme Test		8
4		FH Practice Item		untimed
5		FH Item 1		10
6		FH Item 2		10
7	Questionnaire	Quality Models	Quantity Models	7
8		FH Item 3		10
9	Questionnaire	Quality Models	Quantity Models	7
10		FH Item 4		10
11	Questionnaire	Quality Models	Quantity Models	7
12		FH Item 5		10
13	Questionnaire	Quality Models	Quantity Models	7
14		FH Item 6		10
15	Questionnaire	Quality Models	Quantity Models	7
16		FH Item 7		10
17		Consequences Test		20

Table 2

Intercorrelations[a] and Reliabilities[b] of Independent Variables

(N = 395)

	Sex (M = 1, F = 2)	Test Anxiety	Vocab.	Ideat. Fluency	FH Pretest Score 1	FH Pretest Score 2	FH Pretest Score 3	FH Pretest Score 4	FH Pretest Score 5
Sex	--	.12	.17	.26	.13	.22	.20	.05	.15
Test Anxiety		.87[c]	-.16	.06	-.00	-.03	-.04	-.03	-.03
Vocabulary			.68[d]	.10	.05	.11	.15	.04	.14
Ideational Fluency				.78[d]	.34	.28	.06	.05	.15
FH Pretest Score 1					.54[d]	.79	-.06	.08	-.20
FH Pretest Score 2						.44[d]	.45	.28	.02
FH Pretest Score 3							.54[d]	.53	.38
FH Pretest Score 4								.52[d]	.22
FH Pretest Score 5									.72[d]
Mean	1.5	34.4	14.9	151.3	15.9	11.7	9.3	11.9	39.1
S.D.	.5	16.0	4.7	35.2	5.0	4.5	2.2	2.6	12.4

[a]An R of .10 or greater is significantly different from zero at the 5% level, .13 at the 1% level.

[b]Reliabilities are in the main diagonal.

[c]Reliability is the average interitem correlation corrected for length.

[d]Reliability is the correlation of separately timed halves corrected for double length.

Table 3

Intercorrelations[a] of Dependent Variables for Total Group and for the Three Treatment Groups

	FH Score 1	FH Score 2	FH Score 3	FH Score 4	FH Score 5	Cons. Obvious	Cons. Remote
FH Score 1	.80[b]	.72	-.18	.01	-.23	.28	.34
		.73	-.13	-.05	-.21	.34	.38
		.67	-.16	.03	-.08	.30	.41
		.73	-.21	.04	-.30	.30	.29
FH Score 2		.67	.43	.31	-.07	.22	.19
			.47	.29	-.02	.26	.24
			.49	.40	.10	.23	.20
			.39	.27	-.22	.21	.15
FH Score 3			.60	.53	.33	-.06	-.14
				.57	.36	-.06	-.15
				.59	.40	-.04	-.20
				.41	.21	-.08	-.07
FH Score 4				.48	.14	-.04	-.05
					.18	-.06	-.11
					.15	-.02	-.01
					.09	-.01	-.02
FH Score 5					.87	-.09	-.06
						.10	-.08
						-.18	.02
						-.26	-.12
Cons. Obvious						.85	.15
							.12
							.20
							.10
Cons. Remote							.75
Mean	29.3	18.9	14.4	17.4	62.4	42.8	16.7
	29.4	18.7	14.0	17.3	62.5	43.3	17.0
	27.1	18.0	14.8	17.4	65.6	44.2	16.7
	31.4	19.9	14.3	17.5	59.3	40.9	16.5
S.D.	8.0	5.6	2.7	2.1	19.5	15.1	8.0
	7.9	5.9	2.8	2.2	22.0	14.0	7.6
	6.5	4.5	2.8	2.3	17.9	16.5	8.4
	8.8	6.0	2.5	1.9	17.8	14.7	8.0

[a]The first entry in each cell is the correlation for the total group. The next three entries are the correlations for control, quality, and quantity treatment groups, in that order. N's for the four groups are, respectively, 395, 134, 129, and 132. For the total group, an R of .10 or greater is significantly different from zero at the 5% level, .13 at the 1% level. For the treatment group, an R of approximately .17 is significant at the 5% level, .22 at the 1% level.

[b]Reliabilities (shown in the diagonal) are the average item intercorrelation within the total group corrected for length by the Spearman-Brown formula.

Table 4

Correlations of Independent Variables with Dependent Variables

for Total Group and for the Three Treatment Groups[a]

Independent Variables	Dependent Variables								
	FH Score 1	FH Score 2	FH Score 3	FH Score 4	FH Score 5	Cons. Obvious	Cons. Remote	Mean	S.D.
Sex	.12	.14	.06	.05	.12	.21	-.07	1.5	.5
	.00	.08	.11	.08	.20	.38	-.07	1.5	.5
	.19	.21	.08	-.05	.16	.15	-.07	1.5	.5
	.20	.17	-.01	.13	-.04	.12	-.08	1.5	.5
Test Anxiety	-.03	-.08	-.05	.02	-.05	.05	-.07	34.4	16.0
	-.11	-.06	.09	.20	.01	.06	-.27	32.3	17.4
	.06	-.05	-.15	-.11	-.17	.10	.04	35.2	16.2
	-.01	-.16	-.14	-.09	-.01	.00	.06	35.7	14.0
Vocabulary	.17	.22	.18	.13	.12	.05	.13	14.9	4.7
	.16	.23	.14	.11	.13	.08	.19	14.9	4.7
	.25	.19	.14	.12	.20	.05	.16	14.9	5.0
	.13	.24	.29	.18	.02	.03	.03	15.0	4.5
Ideational Fluency	.32	.23	-.05	-.08	.08	.43	.28	151.3	35.2
	.36	.28	.06	-.06	.15	.47	.29	153.5	35.6
	.32	.23	-.08	-.12	.06	.47	.30	148.1	36.9
	.29	.17	-.12	-.04	.02	.37	.24	152.1	32.9
FH Pretest Score 1	.51	.38	-.10	.03	-.15	.23	.30	15.9	5.0
	.62	.38	-.16	.02	-.13	.26	.35	16.2	5.1
	.43	.39	.01	.08	-.10	.21	.23	16.0	4.9
	.53	.41	-.14	-.01	-.24	.22	.32	15.4	5.1
FH Pretest Score 2	.42	.46	.16	.14	-.02	.18	.20	11.7	4.5
	.48	.41	.03	.12	-.03	.20	.20	11.9	4.3
	.39	.53	.30	.23	.07	.12	.16	11.8	4.6
	.45	.50	.15	.05	-.11	.23	.24	11.4	4.5
FH Pretest Score 3	.04	.30	.45	.23	.27	-.01	-.07	9.3	2.2
	.06	.29	.36	.25	.24	.08	-.07	9.3	1.9
	.09	.41	.54	.33	.35	-.13	-.02	9.3	2.3
	-.01	.24	.45	.11	.26	.05	-.12	9.4	2.4
FH Pretest Score 4	.00	.13	.26	.23	.16	.01	-.01	11.9	2.6
	-.07	.07	.21	.30	.19	.01	.03	12.2	2.4
	-.03	.22	.42	.33	.17	-.09	-.12	11.6	3.0
	.08	.13	.15	.00	.12	.14	.08	11.8	2.5
FH Pretest Score 5	-.17	-.01	.33	.13	.74	-.04	-.04	39.1	12.4
	-.15	.03	.36	.10	.80	.08	.01	39.1	12.2
	-.13	.07	.32	.18	.70	-.13	-.06	39.4	13.1
	-.22	-.11	.30	.10	.74	-.06	-.08	38.7	12.0

[a]The first entry in each cell is the correlation for the total group. The next three entries are the correlations for control, quality, and quantity treatment groups, in that order. N's for the four groups are, respectively, 395, 134, 129, and 132. For the total group, an $\underline{R}$ of .10 or greater is significantly different from zero at the 5% level, .13 at the 1% level. For the treatment groups, an $\underline{R}$ of approximately .17 is significant at the 5% level, .22 at the 1% level.

Table 5

Correlations of Dependent Variables with Canonical Variates

(Covariates: College Attended and Five FH Pretest Scores)

Dependent Variable	Canonical Variate I			Canonical Variate II			Canonical Variate III			Canonical Variate IV		
	TSIA[a]	TSIV	TSVA	TSIA	TSIV	TSVA	TSIA	TSIV	TSVA	TSIA	TSIV	TSVA
FH Score 1	-.21	-.29	-.24	.91	.84	.90	-.14	-.33	-.03	-.04	-.12	-.07
FH Score 2	-.03	-.09	-.08	.83	.88	.76	.46	.30	.55	-.16	.08	-.16
FH Score 3	.34	.34	.32	.13	.31	.02	.88	.84	.88	.17	.19	.20
FH Score 4	.12	.10	.10	.21	.23	.11	.34	.28	.32	.77	.88	.79
FH Score 5	.98	.97	.98	.03	.06	.08	-.19	-.22	-.16	-.02	.01	-.05
Conseq. Obv.	-.09	-.14	-.09	.12	.04	.28	-.05	-.10	-.16	.19	.23	.11
Conseq. Rem.	-.07	-.08	-.08	.40	.35	.48	-.38	-.42	-.34	.13	.10	.15
R	.76	.74	.75	.52	.50	.55	.42	.40	.40	.18	.20	.18
F	13.62	12.71	13.78	7.03	6.43	7.36	4.65	4.23	4.21	1.83	1.85	1.64
p <	.001	.001	.001	.001	.001	.001	.001	.001	.001	.039	.036	.075

[a]These letters designate the design variables used in each of the three MANCOVAs: Treatment, Sex, Anxiety, Ideational Fluency, and Vocabulary.

Table 6

Results of MANCOVA for Treatment, Sex,

Ideational Fluency, and Test Anxiety

Effect	Multivariate significance level	Means		Correlations with canonical variate		p-values for univariate tests
				FH Score 1	.76	.001
	F = 4.377	Control Group	.03	FH Score 5	-.46	.003
Treatment[a]	p < .001	Quality Treatment	-.47	FH Score 2	.46	.003
	R = .369	Quantity Treatment	.44	Cons.-Obv.	-.31	.065
				FH Score 3	-.23	.030
				Cons.-Rem.	-.49	.013
	F = 3.637	Male	-.27	FH Score 2	.45	.023
Sex[b]	p < .001	Female	.27	Cons.-Obv.	.44	.024
	R = .260			FH Score 5	.43	.030
				Cons.-Obv.	.82	.001
Ideational	F =13.654	Low	-.47	FH Score 1	.51	.001
Fluency[b]	p < .001	High	.47	Cons.-Rem.	.48	.001
	R = .462			FH Score 2	.33	.001
				Cons.-Rem.	.57	.033
Test	F = 1.461	Low	.26	FH Score 2	.50	.104
Anxiety[b]	p < .120(n.s.)	Middle	.01	FH Score 5	.36	.259
	R = .220	High	-.27	FH Score 1	.31	.337
Treatment-	F = 1.274					
Test Anxiety	p < .155(n.s.)			FH Score 4	--	.014
Interaction[a]	R = .211					

[a] College attended and the five FH Pretest scores uses as covariates.

[b] College attended used as the only covariate.

Table 7

Results of MANCOVA for Treatment, Sex,

Ideational Fluency, and Vocabulary

Effect	Multivariate significance level	Means		Correlations with canonical variate		p-values for univariate tests
				FH Score 1	.74	.001
	F = 4.455	Control Group	.04	FH Score 2	.45	.003
Treatment[a]	p < .001	Quality Treatment	-.46	FH Score 5	-.44	.004
	R = .371	Quantity Treatment	.42	Cons.-Obv.	-.32	.061
				FH Score 3	-.22	.029
				Cons.-Rem.	-.54	.009
	F = 3.286	Male	-.24	Cons.-Obv.	.53	.011
Sex[b]	p < .002	Female	.24	FH Score 2	.35	.094
	R = .248			FH Score 5	.33	.115
				Cons.-Obv.	.81	.001
Ideational	F = 8.459	Low	-.65	FH Score 1	.51	.001
Fluency[b]	p < .001	Middle	-.05	Cons.-Rem.	.46	.001
	R = .497	High	.70	FH Score 2	.30	.005
				FH Score 3	.80	.001
	F = 4.827	Low	-.28	FH Score 4	.66	.001
Vocabulary[b]	p < .001	High	.28	FH Score 2	.58	.001
	R = .296			FH Score 5	.51	.003
				FH Score 2	.64	.032
Sex-	F = 1.598			FH Score 4	.61	.041
Vocabulary	p < .135(n.s.)			FH Score 3	.60	.043
Interaction[b]	R = .176			Cons.-Rem.	.45	.132
				Cons.-Obv.	-.33	.271

[a] College attended and the five FH Pretest scores used as covariates.

[b] College attended used as the only covariate.

Table 8

Results of MANCOVA for Treatment, Sex,

Vocabulary, and Test Anxiety

Effect	Multivariate significance level	Means		Correlations with canonical variate		p-values for univariate tests
				FH Score 1	.80	.001
	F = 4.162	Control Group	.01	FH Score 2	.48	.002
Treatment[a]	p < .001	Quality Treatment	-.46	FH Score 5	-.48	.002
	R = .363	Quantity Treatment	.45	FH Score 3	-.26	.027
				Cons.-Obv.	.67	.001
	F = 4.742	Male	-.30	FH Score 2	.46	.008
Sex[b]	p < .001	Female	.30	FH Score 1	.40	.020
	R = .294			FH Score 5	.34	.049
				FH Score 3	.78	.001
	F = 5.128	Low	-.31	FH Score 4	.65	.001
Vocabulary[b]	p < .001	High	.31	FH Score 2	.58	.001
	R = .304			FH Score 5	.46	.005
				Cons.-Rem.	.59	.032
Test	F = 1.415	Low	.25	FH Score 2	.42	.219
Anxiety[b]	p < .140	Middle	-.02	FH Score 1	.32	.324
	R = .214	High	-.23	FH Score 4	-.30	.336
				Cons.-Obv.	-.53	.026
Sex-	F = 2.498			FH Score 2	.52	.029
Vocabulary	p < .016			FH Score 3	.51	.033
Interaction[b]	R = .218			FH Score 4	.50	.037
Treatment-	F = 1.261					
Test Anxiety	p < .164(n.s.)			FH Score 4	--	.012
Interaction[a]	R = .216					

[a] College attended and the five FH Pretest scores used as covariates.

[b] College attended used as the only covariate.